Abstractive text summarization datasets, models, and tokenization approaches for Turkish and Hungarian

Batuhan Baykara and **Tunga Güngör**
Boğaziçi University, Computer Engineering, Istanbul, Turkey

Relevant UniDive working groups: WG1, WG3

1 Introduction

Text summarization is the process of automatically generating brief, fluent, and salient text from a document (Edmundson (1969); Luhn (1958); Nenkova and McKeown (2012)). Summarization can be divided into two as abstractive and extractive (Hahn and Mani (2000)). Recent advances in deep learning enabled significant progress in natural language understanding and generation tasks, including abstractive summarization. Despite such advances, the works are mostly limited to English which prevents progress in resource-scarce languages.

Agglutinative languages such as Turkish and Hungarian differ from other languages in the sense that the word formation process heavily depends on affixation. The morpho-syntactic properties of these languages enable the word to carry more information and utilizing morphology was shown to be effective for tasks such as named entity recognition (Güngör et al. (2019)), part-of-speech tagging (Eşref and Can (2019)), learning word embeddings (Dobrossy et al. (2019); Üstün et al. (2018)), and machine translation (Pan et al. (2020)).

In this work, we build text summarization datasets and research on different abstractive summarization models for agglutinative languages. The contributions are as follows:

- We release two large-scale publicly available summarization datasets for low-resource agglutinative languages Turkish and Hungarian.
- We provide strong baselines for both datasets.
- Two types of morphological tokenization methods (SeparateSuffix and CombinedSuffix) are proposed for both Turkish and Hungarian. Through these methods, the effect of morphology is studied on both datasets.
- We use the pointer-generator model as a baseline that is commonly-used in abstractive text summarization and compare it with a state-of-the-art BERT-based approach.

2 Related Work

Turkish text summarization approaches have been mostly limited to extractive methods. Studies made use of latent semantic analysis and singular value decomposition (Özsoy et al. (2010)), similarity and frequency based metrics (Çığır et al. (2009)), non-negative matrix factorization (Güran et al. (2011)), semantic information (Güran et al.), and query based models (Pembe and Güngör (2008)). The datasets used in all these studies are highly limited in size ranging from 50 (Özsoy et al. (2010)) to 120 (Çığır et al. (2009)) documents. Hungarian text summarization has been studied even less than Turkish. It has been employed on speech data using traditional scoring methods (Beke and Szaszák (2016)) or for analyzing error propagation in speech summarization (Ákos Tündik et al. (2019)).

3 Datasets

The sizes of text summarization datasets are critical for abstractive summarization where mostly deep learning-based approaches are utilized. In this work, we prepare two large-scale datasets, TR-News for Turkish and HU-News for Hungarian. Both datasets were compiled in a manner to make them suitable for other NLP tasks such as topic classification, author identification, and headline generation. An approach similar to the one used in the compilation of the English CNN/Daily Mail dataset was adopted. First, all publicly available newspapers for the two languages were gathered from Wikipedia. By a detailed analysis based on criteria such as content and abstract lengths, HTML markup quality, content quality, three news sites for both Turkish and Hungarian were identified. A web crawler was used to extract the relevant fields which are URL, title, abstract, content, date of publish, author, source, topic, and tags. The documents were further processed to eliminate the ones with missing values in content or abstract fields.

The training, validation, and test sets are, respectively, 277,573, 14,610, and 15,379 for TR-News, and 211,860, 11,151, and 11,738 for HU-News.

	TR-News			HU-News		
Model	R1	R2	RL	R1	R2	RL
LEAD-2	31.37	17.91	26.92	24.34	7.87	17.61
LEAD-3	28.64	16.21	24.07	23.70	7.78	16.75
WhiteSpace	31.61	18.55	29.57	22.92	7.69	19.78
Unigram LM	33.38	19.77	31.15	**24.33**	**8.25**	**20.91**
SeparateSuffix	**34.94**	**20.89**	**32.56**	23.86	8.10	20.53
CombinedSuffix	33.93	20.07	31.57	23.57	7.97	20.23
mBERT-uncased	21.70	8.95	18.41	21.88	4.51	17.62
mBERT-cased	**30.99**	**18.09**	**26.54**	**26.54**	9.72	**19.51**
BERTurk-uncased-32K	27.40	15.60	23.36	-	-	-
BERT-uncased-128K	26.92	15.25	22.96	-	-	-
huBERT-uncased	-	-	-	25.40	**10.03**	18.54

Table 1: Rouge-1, Rouge-2, and Rouge-L results of pointer-generator models with different tokenizations and BERT models.

4 Methodology

Two models have been used in this study for text summarization, which are the pointer-generator model (See et al. (2017)) and the BERT+Transformer model. As the first and the baseline model, we chose the pointer-generator model which is commonly-used in abstractive summarization. It is an encoder-decoder network based on the LSTM architecture and is capable of deciding whether to point to a word from the input sequence or to generate a new word from the vocabulary at each time step. As the second model, we utilized an encoder-decoder architecture that makes use of BERT as the encoder and a 6-layered transformer network as the decoder (Liu and Lapata (2019)). To initialize the encoder, we used a pretrained BERT (BERTSumAbs) model.

To see the effect of morphology-based tokenization in abstractive summarization for agglutinative languages, we implemented two different tokenizers for Turkish and Hungarian. The approaches we use are more linguistically-oriented compared to the commonly-used unigram language model (ULM) and byte pair encoding (BPE) tokenizations. Rather than splitting the word based on statistical methods, we aim to leverage the true morphological structure within the words. Both methods are based on the roots of the words and the suffixes. In the first method (SeparateSuffix) all morphemes (root and suffixes) are considered separately, whereas in the second one (CombinedSuffix) the word is divided into two parts as the root and all the suffixes in concatenated form.

5 Experiments and Results

In the first experiment, we test the effects of different tokenization methods using the pointer-generator model. In the original model, whitespace tokenization is used. In this experiment, in addition to whitespace that serves as a baseline, we use three other tokenization methods. Two of them (SeparateSuffix and CombinedSuffix) are linguistically-oriented and one (ULM) is statistical. The first part in Table 1 shows the LEAD baselines that are commonly-used in text summarization and considered as strong baselines, and the second part shows the tokenization results. The results show that morphological tokenization methods are effective for both agglutinative languages compared to whitespace tokenization. When we compare the two morphology-based methods, we see that SeparateSuffix outperforms CombinedSuffix.

The second experiment aims at observing the performance of a state-of-the-art summarization model and comparing its performance with the baseline pointer-generator model. In addition to using multilingual BERT models, we also experiment with the monolingual BERT models which are BERTurk (Schweter (2020)) for Turkish and huBERT (Nemeskey (2020)) for Hungarian. The third part in the table shows the results. We see that the multilingual cased BERT model outperforms all the other BERT models for both Turkish and Hungarian. The best BERT models for Hungarian outperform both of the LEAD baselines and the pointer-generator models. This is not the case for Turkish where the best BERT model falls behind the pointer-generator model.

References

András Beke and György Szaszák. 2016. Automatic summarization of highly spontaneous speech. In *Proceedings of Speech and Computer - SPECOM*, pages 140–147.

Balint Dobrossy, Márton Makrai, Balázs Tarján, and György Szaszák. 2019. Investigating sub-word embedding strategies for the morphologically rich and free phrase-order hungarian. In *Proceedings of the 4th Workshop on Representation Learning for NLP*, pages 187–193.

Harold P. Edmundson. 1969. New methods in automatic extracting. *Journal of the Association for Computing Machinery*, 16(2):264–285.

Yasin Eşref and Burcu Can. 2019. Scalable training of L1-regularized log-linear models. In *Proceedings of the 27th Signal Processing and Communications Applications Conference*, pages 1–4.

Onur Güngör, Tunga Güngör, and Suzan Üsküdarlı. 2019. The effect of morphology in named entity recognition with sequence tagging. *Natural Language Engineering*, 25(1):147–169.

Aysun Güran, Nilgün Güler Bayazıt, and Eren Bekar. 2011. Automatic summarization of turkish documents using non-negative matrix factorization. In *Proceedings of the International Symposium on Innovations in Intelligent Systems and Applications*, pages 480–484.

Aysun Güran, Nilgün Güler Bayazıt, and Mustafa Zahid Gürbüz. Efficient feature integration with wikipedia-based semantic feature extraction for turkish text summarization. *Turkish Journal of Electrical Engineering and Computer Sciences*, 21(5):1411–1425.

Udo Hahn and Inderjeet Mani. 2000. The challenges of automatic summarization. *Computer*, 33(11):29–36.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3730–3740.

H. P. Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

Dávid Márk Nemeskey. 2020. Egy embert próbáló feladat. In *Proceedings of the 16th Magyar Számítógépes Nyelvészeti Konferencia*, pages 409–418.

Ani Nenkova and Kathleen McKeown. 2012. A survey of text summarization techniques. *In: Mining Text Data*, pages 43–76.

Yirong Pan, Xiao Li, Yating Yang, and Rui Dong. 2020. Morphological word segmentation on agglutinative languages for neural machine translation. arXiv:2001.01589.

Fatma Canan Pembe and Tunga Güngör. 2008. Towards a new summarization approach for search engine results: An application for turkish. In *Proceedings of the 23rd International Symposium on Computer and Information Sciences*, pages 1–6.

Stefan Schweter. 2020. Berturk - bert models for turkish.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1073–1083.

Máté Ákos Tündik, Valér Kaszás, and György Szaszák. 2019. Assessing the semantic space bias caused by asr error propagation and its effect on spoken document summarization. In *Proceedings of the 20th Interspeech*, pages 1333–1337.

Celal Çığır, Mücahid Kutlu, and İlyas Çiçekli. 2009. Generic text summarization for turkish. In *Proceedings of the 24th International Symposium on Computer and Information Sciences*, pages 224–229.

Makbule Özsoy, İlyas Çiçekli, and Ferda Alpaslan. 2010. Text summarization of turkish texts using latent semantic analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 869–876.

Ahmet Üstün, Murathan Kurfalı, and Burcu Can. 2018. Characters or morphemes: How to represent words? In *Proceedings of the 3rd Workshop on Representation Learning for NLP*, pages 144–153.